

# Managing Confidentiality and Provenance across Mixed Private and Publicly-Accessed Data and Metadata

Lars Vilhuber<sup>1</sup> John M. Abowd<sup>1</sup> William Block<sup>2</sup>  
Carl Lagoze<sup>3</sup> Jeremy Williams<sup>2</sup>

<sup>1</sup>Labor Dynamics Institute, ILR,

<sup>2</sup>Cornell Institute for Social and Economic Research,

<sup>3</sup>University of Michigan

November 2013, FCSM 2013

# Introduction

## NCRN

- ▶ This work is part of the NSF Census Research Network (NCRN) - Cornell Node ("Integrated Research Support, Training and Data Documentation")
- ▶ Funded by NSF Grant #1131848.
- ▶ For more information, see [www.ncrn.cornell.edu](http://www.ncrn.cornell.edu).



# Introduction

## Overview of work

- ▶ Basic program outlined in Abowd, Vilhuber, and Block (PSD 2012) [3] and Lagoze, Block, Williams, Abowd, and Vilhuber, (IDCC 2013) [8]
- ▶ PROV extension described in more detail in Lagoze, Williams, Vilhuber (Metadata and Semantics Research Conference, November 2013) and Lagoze et al (European DDI User Confernce, December 2013) [9]

## Introduction

### Some facts that motivated us

#### Stating the problem in the U.S. case

#### CED<sup>2</sup>AR: A proposed solution

- What is DDI

- DDI extension for confidentiality protection

- DDI extension for provenance tracing

# Replication of research results

## Critical element of science

- ▶ Replication of methods, data inputs, computational environment is a critical element of the scientific approach
- ▶ Journals, funding agencies (in the U.S.) have been moving to making archiving of inputs to scientific results more robust, even mandatory

# Not a new problem

## Econometrica

“In its first issue, the editor of Econometrica (1933), Ragnar Frisch, noted the importance of publishing data such that readers could fully explore empirical results. Publication of data, however, was discontinued early in the journal’s history. [...] The journal arrived full-circle in late 2004 when Econometrica adopted one of the more stringent policies on availability of data and programs.

<http://www.econometricsociety.org/submissions.asp#4> as cited in Anderson et al (2005)

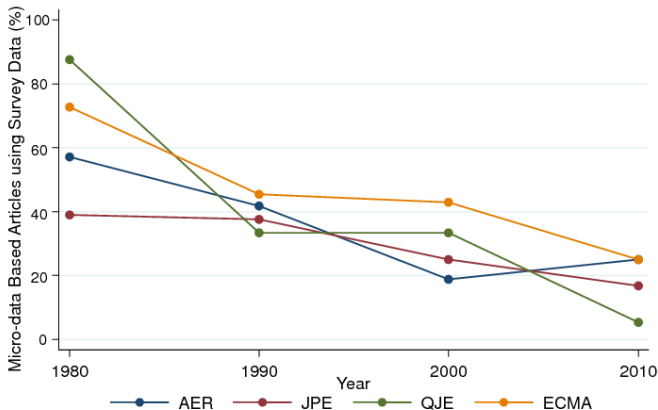
# Problem will become worse

## Increased use of restricted-access data

- ▶ Today's young scholars pursue research programs that mandate inherently identifiable data
  - ▶ Geospatial relations,
  - ▶ Exact genome data,
  - ▶ Networks of all sorts,
  - ▶ Linked administrative records
- ▶ These researchers acquire authorized, generally unfettered, restricted access to the confidential, identifiable data and perform their analyses in secure environments.
- ▶ Archiving (curation) of input data is complicated
- ▶ Knowledge discovery is complicated

# Decline in the use of classic public-use data

Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010

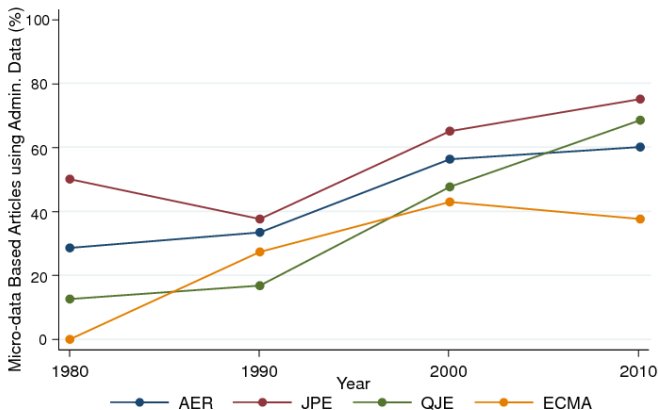


Note: "Pre-existing survey" datasets refer to micro surveys such as the CPS, BHPS and so on, include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.



# Increase in the use of administrative data in economics

Use of Administrative Data in Publications in Leading Journals, 1980-2010



Note: "Administrative" data is data collected on individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

# Not limited to economics

## Nature, 2012

“Many of the emerging ‘big data’ applications come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results.”

(Huberman, Nature 482, 308 (16 February 2012) doi:10.1038/482308d)

## Other domains

- ▶ Biology (genetics data, chemical compounds)
- ▶ Computer science (search records, single-firm examples)

## Introduction

Some facts that motivated us

Stating the problem in the U.S. case

CED<sup>2</sup>AR: A proposed solution

- What is DDI

- DDI extension for confidentiality protection

- DDI extension for provenance tracing

# Why we think there is a problem

## Core issues

- a Insufficient curation (starting with archiving)
- b No way to reference data (unique identifiers)
- c No consistent way to learn about the data (metadata dissemination)
- d Weak or non-existent provenance tracing

# Generalized problem

## Multiple data sources in the US

- ▶ U.S. Census Bureau (RDC) ▶ more
- ▶ Internal Revenue Service (confidential, public-use) ▶ more
- ▶ Bureau of Labor Statistics (confidential, public-use data) ▶ more

## Present elsewhere?

- ▶ Canada:
  - ▶ Centre for Data Development and Economic Research (CDER: RDC-like for business data) ▶ more
  - ▶ better: Canadian RDC network ▶ more
- ▶ France: Réseau Quetelet ▶ more , Centre d'accès sécurisé distant aux données (CASD)
- ▶ Germany: IAB

## Introduction

Some facts that motivated us

Stating the problem in the U.S. case

## CED<sup>2</sup>AR: A proposed solution

What is DDI

DDI extension for confidentiality protection

DDI extension for provenance tracing

# Comprehensive Extensible Data Documentation and Access (CED<sup>2</sup>AR)

## Core

We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols, and allows for much improved provenance tracing.

# Proposed solution

## Extensible framework

- ▶ Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms and provenance tracing



# Proposed solution

## Extensible framework

- ▶ Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms and provenance tracing
- ▶ Connectors (import/export) to other sources and standards

# Proposed solution

## Extensible framework

- ▶ Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms and provenance tracing
- ▶ Connectors (import/export) to other sources and standards
- ▶ To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)

# Proposed solution

## Extensible framework

- ▶ Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms and provenance tracing
- ▶ Connectors (import/export) to other sources and standards
- ▶ To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)
- ▶ Interim solution for those datasets without unique identifiers (Digital Object Identifier, DOI)

# Proposed solution

## Extensible framework

- ▶ Based on existing standards (Data Documentation Initiative, DDI) with **extension to accomodate disclosure protection mechanisms and provenance tracing**
- ▶ Connectors (import/export) to other sources and standards
- ▶ To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)
- ▶ Interim solution for those datasets without unique identifiers (Digital Object Identifier, DOI)

## What is DDI?

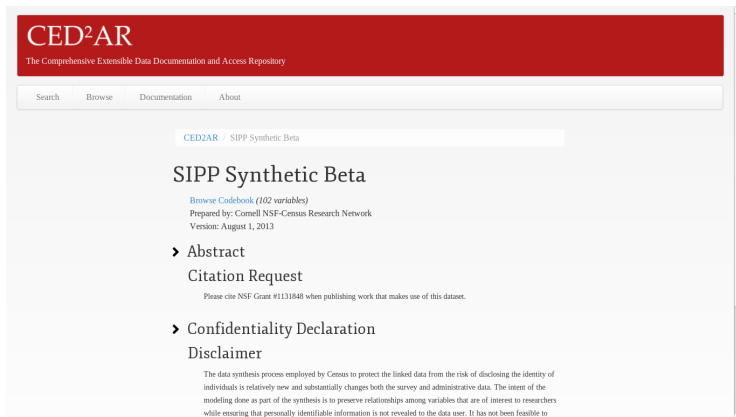
# Example of DDI

```

<?xml version="1.0" encoding="UTF-8"?>
<codeBook xmlns="ddi:codebook:2_5" ...>
  <docDscr>
    <citation>
      <titlStmt>
        <titl>SIPP_Synthetic_Beta</titl>
        <altTitl>SSB</altTitl>
        <IDNo agency="DOI">TBD</IDNo>
      </titlStmt>
      <rspStmt>
        <AuthEnty affiliation="Cornell University">
          Virtual RDC
        </AuthEnty>
      </rspStmt>
    </citation>
  </docDscr>
</codeBook>

```

# ..better seen as



The screenshot shows the CED2AR website interface. At the top is a red header with the text "CED<sup>2</sup>AR" and "The Comprehensive Extensible Data Documentation and Access Repository". Below this is a navigation bar with links for "Search", "Browse", "Documentation", and "About". The main content area has a breadcrumb trail "CED2AR / SIPP Synthetic Beta". The title "SIPP Synthetic Beta" is prominently displayed, followed by a link to "Browse Codebook (102 variables)". Below the title, it states "Prepared by: Cornell NSF-Census Research Network" and "Version: August 1, 2013". There are two main sections: "Abstract" and "Confidentiality Declaration", each with a "Citation Request" link. The "Confidentiality Declaration" section includes a "Disclaimer" paragraph explaining the data synthesis process and its purpose to protect individual identities while preserving relationships between variables.

**CED<sup>2</sup>AR**  
The Comprehensive Extensible Data Documentation and Access Repository

Search Browse Documentation About

CED2AR / SIPP Synthetic Beta

## SIPP Synthetic Beta

[Browse Codebook \(102 variables\)](#)  
Prepared by: Cornell NSF-Census Research Network  
Version: August 1, 2013

- ▶ **Abstract**  
[Citation Request](#)  
Please cite NSF Grant #1131848 when publishing work that makes use of this dataset.
- ▶ **Confidentiality Declaration**  
[Disclaimer](#)  
The data synthesis process employed by Census to protect the linked data from the risk of disclosing the identity of individuals is relatively new and substantially changes both the survey and administrative data. The intent of the modeling done as part of the synthesis is to preserve relationships among variables that are of interest to researchers while ensuring that personally identifiable information is not revealed to the data user. It has not been feasible to

# Example DDI: ICPSR

[Log In/Create Account](#)

[Find & Analyze Data](#)
[Membership in ICPSR](#)
[Deposit Data](#)
[ICPSR Summer Program](#)
[Resources for Instructors](#)
[Data Management & Curation](#)

**ICPSR** Find & Analyze Data

[Find Data](#)
[Search/Compare Variables](#)
[Find Publications](#)
[Resources for Students](#)
[Get Help](#)

**Table of Contents**

- [Top of page](#)
- [Access Notes](#)
- [Dataset\(s\)](#)
- [Study Description](#)
  - [Citation](#)
  - [Scope of Study](#)
  - [Methodology](#)
  - [Version\(s\)](#)
- [Related Publications](#)
- [Variables](#)
- [Utilities](#)
- [Metadata Exports](#)
- [Download Statistics](#)

series: Survey of Income and Program Participation (SIPP) Series > study: Survey of Income and Program Participation (SIPP) 2004 Panel >

[<< Back to results](#)

Result 1 of 369 >>

## Survey of Income and Program Participation (SIPP) 2004 Panel (ICPSR 4517)

**Principal Investigator(s):** United States Department of Commerce. Bureau of the Census

**Summary:**  
 This data collection is part of a longitudinal survey designed to provide detailed information on the economic situation of households and persons in the United States. These data examine the distribution of income, wealth, and poverty in American society and gauge the effects of federal and state programs on the well-being of families and individuals. There are three basic elements contained in the survey. The first is a control card that records basic social and demographic characteristics for... [\(more info\)](#)

**Series:** [Survey of Income and Program Participation \(SIPP\) Series](#)

**Access Notes**

- These data are freely available.

**Dataset(s)**

**WARNING:** Because this study has many datasets, the **download all files** option has been suppressed, and you will need to download one dataset at a time.


**WARNING:** This study is over 150MB in size and may take several minutes to download on a typical internet connection.

**Child Care & Early Education Research Connections**

This study is provided by [Child Care & Early Education Research Connections](#).



# Example DDI: UK data archive

UK Data Service


**Discover**

Variable and question bank

[Site Search](#)
[FAQ](#)
[Help](#)
[Contact](#)

[About us](#)
[Get data](#)
[Use data](#)
[Manage data](#)
[Deposit data](#)
[News and Events](#)

Discover > Series

## Series

UK Data Service series record for:

### Family Expenditure Survey

[Abstract](#) | [Access](#) | [Related](#) | [Search](#)

#### SERIES ABSTRACT

The Family Expenditure Survey (FES), which ran from 1961-2001, was a continuous annual survey that provided information on household and personal incomes, certain payments that recurred regularly (e.g. rent, gas and electricity bills, telephone accounts, insurances, season tickets and hire purchase payments), and included a detailed 14-day expenditure record. From 2001, the both the FES and the National Food Survey (NFS) were replaced by a new survey, the Expenditure and Food Survey (EFS), which subsequently became the Living Costs and Food Survey (LCF) from 2008.

#### DATA ACCESS

GN 33057 | FAMILY EXPENDITURE SURVEY, 1961-2001

#### RELATED RESOURCES

**Related studies:**

Family Resources Survey, 1979 (SN 1930)

Family Expenditure Survey Follow-up Survey of Disabled Adults, 1986-1987 (SN 2943)

Institute for Fiscal Studies Households Below Average Income Dataset, 1961-1991 (SN 3300)

Simulation Program for Indirect Taxation, 1988: SPIT Version 6 (SN 3328)

Imputed Expenditure Variables for Family Resources Survey, 1995-1996 (SN 4407)

Living Costs and Food Survey, 2006-2011: Secure Access (SN 7047)

# Expanded DDI attributes

## Standard DDI

### Fragment of variable description\*

```

<var ID="V1" dcml="0" files="F1" intrvl="discrete"
  name="cur_end mar_flag">
  <location width="12"/>
    <labl>Flag: Linked marriage ended</labl>
    <valrng>
      <range UNITS="REAL" max="2" min="0"/>
    </valrng>
    <sumStat type="vald"> 123 </sumStat>
    <sumStat type="invd"> 456 </sumStat>
    <catgry>
      <catValu> 1 </catValu>
      <catStat type="freq"> 234 </catStat>
    </catgry>
  </var>

```

\* All values are fake

# Expanded DDI attributes

## Standard DDI

### Fragment of variable description\*

```
<!--var ID="V1" dcml="0" files="F1" intrvl="discrete"
name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl -->
    <valrng>
      <range UNITS="REAL" max="2" min="0"/>
    </valrng>
    <!-- sumStat type="vald"> 123 </sumStat>
    <sumStat type="invd"> 456 </sumStat>
    <catgry>
      <catValu> 1 </catValu>
      <catStat type="freq"> 234 </catStat>
    </catgry -->
```

\* All values are fake

# Expanded DDI attributes

## Standard DDI

### Fragment of variable description\*

```
<!--var ID="V1" dcml="0" files="F1" intrvl="discrete"
name="cur_end mar_flag">
  <location width="12"/>
    <labl>Flag: Linked marriage ended</labl>
    <valrng>
      <range UNITS="REAL" max="2" min="0"/>
    </valrng -->
    <sumStat type="vald"> 123 </sumStat>
    <sumStat type="invd"> 456 </sumStat>
  <!-- catgry>
    <catValu> 1 </catValu>
    <catStat type="freq"> 234 </catStat>
  </catgry -->
```

\* All values are fake

# Expanded DDI attributes

## Enhanced DDI

Re-using existing attribute, but expanding scope.\*

```
<var ID="V1" dcml="0" files="F1" intrvl="discrete"
  name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng access="release">
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
  <sumStat access="restricted" type="vald"> 123 </sumStat>
  <sumStat access="restricted" type="invd"> 456 </sumStat>
  <catgry access="release">
    <catValu access="release"> 1 </catValu>
    <catStat type="freq" access="restricted">
      234
    </catStat>
  </catgry>
```

\* All values are fake

# Expanded DDI attributes

## Enhanced DDI

Allows for verifiable filtering\*

```
<var ID="V1" dcml="0" files="F1" intrvl="discrete"
  name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng access="release">
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
  <!-- sumStat suppressed -->
  <!-- sumStat suppressed -->
  <catgry access="release">
    <catValu access="release"> 1 </catValu>
    <catStat type="freq" access="restricted">
      [suppressed]
    </catStat>
  </catgry>
```

\* All values are fake

# Application to confidentiality protection

## Browse all Variables

Searching Synthetic Longitudinal Business Database Uncheck any variables to be released, then press **[SAVE]**.

Show  variables

Confidential	Variable Name	Label	Codebook
<input checked="" type="checkbox"/>	act	dropped Activity Code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestnaics	dropped Best NAICS code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestsic	dropped Best SIC code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cbp	dropped	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cfn	dropped Census File Number	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	county	masked County FIPS codes	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	emp	synthetic March 12 Employment	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	firstflag	dropped First Link Flag	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	firstyear	synthetic First Year Establishment is Observed	Synthetic Longitudinal Business Database

# Options

- ▶ Variable is suppressed, including all subordinate elements
- ▶ Variable description is released, but all subordinate statistical elements are suppressed (attribute of `< var >` set to "released") [default]
- ▶ Expand all existing attributes, individually select subordinate elements to suppress (attribute of sub-element is set to "suppressed", content suppressed)



# Application to confidentiality protection

## Browse all Variables

Searching Synthetic Longitudinal Business Database Uncheck any variables to be released, then press **[SAVE]**.

Show  variables

Confidential	Variable Name	Label	Codebook
<input checked="" type="checkbox"/>	act	dropped Activity Code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestnaics	dropped Best NAICS code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestsic	dropped Best SIC code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cbp	dropped	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cfm	dropped Census File Number	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	county	masked County FIPS codes	Synthetic Longitudinal Business Database
<input type="checkbox"/>	emp	synthetic March 12 Employment	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	firstflag	dropped First Link Flag	Synthetic Longitudinal Business Database
<input type="checkbox"/>	firstyear	synthetic First Year Establishment is Observed	Synthetic Longitudinal Business Database

# Application to confidentiality protection

## Browse all Variables

Searching Synthetic Longitudinal Business Database Uncheck any variables to be released, then press **[SAVE]**.

Show  variables

Confidential	Variable Name	Label	Codebook
<input checked="" type="checkbox"/>	act	dropped Activity Code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestnaics	dropped Best NAICS code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	bestsic	dropped Best SIC code	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cbp	dropped	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	cfm	dropped Census File Number	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	county	masked County FIPS codes	Synthetic Longitudinal Business Database
<input type="checkbox"/>	emp	synthetic March 12 Employment	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/>	firstflag	dropped First Link Flag	Synthetic Longitudinal Business Database
<input type="checkbox"/>	firstyear	synthetic First Year Establishment is Observed	Synthetic Longitudinal Business Database

# Implementation

## Definitions

- ▶ First draft of specification in test use by our team

# Implementation

## Definitions

- ▶ First draft of specification in test use by our team
- ▶ Full enhanced specification (based on DDI-Codebook 2.5) published on CED<sup>2</sup>AR

# Implementation

## Definitions

- ▶ First draft of specification in test use by our team
- ▶ Full enhanced specification (based on DDI-Codebook 2.5) published on CED<sup>2</sup>AR
- ▶ Enhanced specification proposed to DDI Alliance

# Implementation

## Definitions

- ▶ First draft of specification in test use by our team
- ▶ Full enhanced specification (based on DDI-Codebook 2.5) published on CED<sup>2</sup>AR
- ▶ Enhanced specification proposed to DDI Alliance
- ▶ Expand to DDI-Lifecycle

# Provenance

## The provenance problem

“data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources” [...] “from it, one can ascertain the quality of the data base and its ancestral data and derivations, track back sources of errors, allow automated reenactment of derivations to update the data, and provide attribution of data sources”

Simmhan, Plale, and Gannon, “A survey of data provenance in e-science,” ACM Sigmod Record, 2005

# Support in DDI

## Provenance and Metadata

Not (currently) a “native” component of DDI, closest thing is:


```
<xs:complexType name="othrStdyMatType">
  <xs:complexContent>
    <xs:extension base="baseElementType">
      <xs:sequence>
        <xs:element ref="relMat" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="relStdy" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="relPubl" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="othRefs" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

## Downside

No structure. Mostly verbose entries.



# Only a verbose description

UK Data Service


Discover
Variable and question bank

Site Search
FAQ
Help
Contact

About us
Get data
Use data
Manage data
Deposit data
News and Events

Discover > Series

## Series

UK Data Service series record for:

### Family Expenditure Survey

[Abstract](#)
[Access](#)
[Related](#)
[Search](#)

---

#### SERIES ABSTRACT

The Family Expenditure Survey (FES), which ran from 1961-2001, was a continuous annual survey that provided information on household and personal incomes, certain payments that occurred regularly (e.g. rent, gas and electricity bills, telephone accounts, insurances, season tickets and hire purchase payments), and included a detailed 14-day expenditure record. From 2001, the both the FES and the National Food Survey (NFS) were replaced by a new survey, the Expenditure and Food Survey (EFS), which subsequently became the Living Costs and Food Survey (LCF) from 2008.

---

#### DATA ACCESS

GN 33057 | FAMILY EXPENDITURE SURVEY, 1961-2001

---

#### RELATED RESOURCES

**Related studies:**

Family Resources Survey, 1979 (SN 1930)

Family Expenditure Survey Follow-up Survey of Disabled Adults, 1986-1987 (SN 2943)

Institute for Fiscal Studies Households Below Average Income Dataset, 1961-1991 (SN 3300)

Simulation Program for Indirect Taxation, 1988; SPIT Version 6 (SN 3328)

Imputed Expenditure Variables for Family Resources Survey, 1995-1996 (SN 4407)

Living Costs and Food Survey, 2006-2011: Secure Access (SN 7047)

# UK Data Archive example

[Abstract](#) | [Access](#) | [Related](#) | [Search](#)

## SERIES ABSTRACT

The Family Expenditure Survey (FES), which ran from 1961-2001, was a continuous annual survey that provided information on household and personal incomes, certain payments that recurred regularly (e.g. rent, gas and electricity bills, telephone accounts, insurances, season tickets and hire purchase payments), and included a detailed 14-day expenditure record. From 2001, the both the FES and the National Food Survey (NFS) were replaced by a new survey, the Expenditure and Food Survey (EFS), which subsequently became the Living Costs and Food Survey (LCF) from 2008.

## DATA ACCESS

GN 33057 | FAMILY EXPENDITURE SURVEY, 1961-2001



## RELATED RESOURCES

### Related studies:

Family Resources Survey, 1979 (SN 1930)

# Provenance (cont)

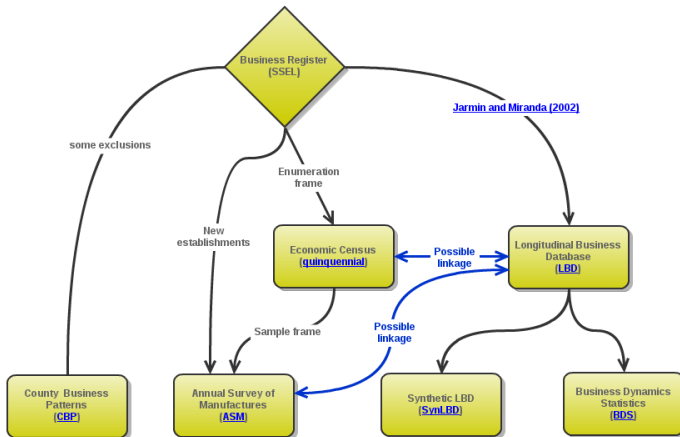
## PROV model

W3C PROV Model based in the notions of

1. **entities** that are physical, digital, and conceptual things in the world;
2. **activities** that are dynamic aspects of the world that change and create entities; and
3. **agents** that are responsible for activities.
4. a set of **relationships** that can exist between them that express attribution, delegation, derivation, etc.

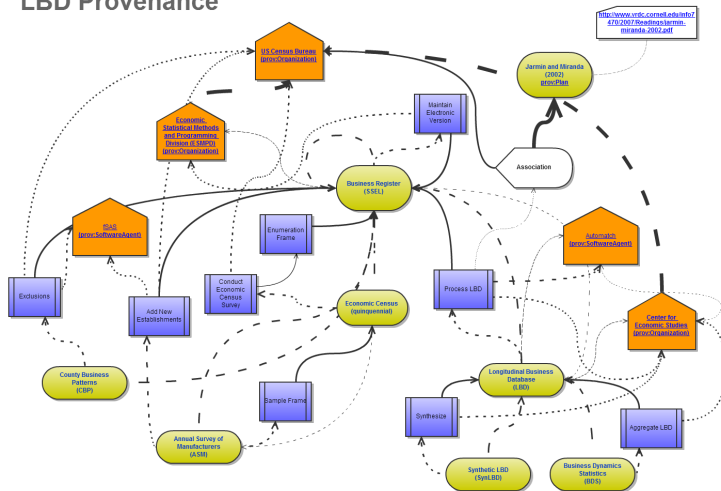
# Incorporating PROV (LBD)

## LBD Provenance

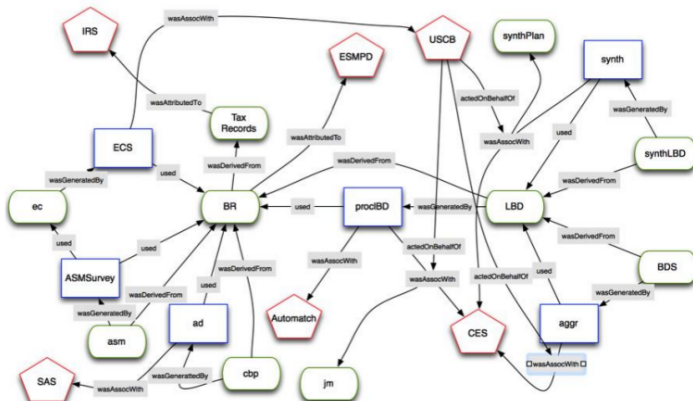


# Incorporating PROV (LBD)

## LBD Provenance



# Incorporating PROV (LBD)



# PROV as RDF

```

entity(cdr:LBD, [prov:type='cdr:dataset', prov:label="Longitudinal Business Data"])
entity(cdr:synthLBD, [prov:type='cdr:dataset', prov:label="Synthetic LBD"])
entity(cdr:BDS, [prov:type='cdr:dataset', prov:label="Business Dynamics Statistics"])
entity(cdr:BR, [prov:type='cdr:dataset', prov:label="Business Register"])
entity(cdr:cbp, [prov:type='cdr:dataset', prov:label="County Business Patterns"])
entity(cdr:asm, [prov:type='cdr:dataset', prov:label="Annual Survey of Manufacturers"])
entity(cdr:ec, [prov:type='cdr:dataset', prov:label="Economic Census"])
entity(cdr:jm, [prov:type='prov:Plan', prov:label="Jarmin Miranda 2002"])
entity(cdr:synthPlan, [prov:type='prov:Plan', prov:label="synthetic plan"])
entity(cdr:tax, [prov:type='cdr:dataSet', prov:label="IRS Tax Records"])

agent(cdr:USCB, [prov:type='prov:Organization', prov:label="US Census Bureau"])
agent(cdr:CES, [prov:type='prov:Organization', prov:label="Center for Economic Studies"])
agent(cdr:IRS, [prov:type='prov:Organization', prov:label="Internal Revenue Service"])
agent(cdr:autoMatch, [prov:type='prov:SoftwareAgent'])
agent(cdr:SAS, [prov:type='prov:SoftwareAgent'])
agent(cdr:ESMPD, [prov:type='prov:SoftwareAgent',
  prov:label="Economic Statistical Methods and Programming Division"])

activity(cdr:synth, [prov:label="anonymize"])
activity(cdr:aggr, [prov:label="aggregate"])
activity(cdr:procLBD, [prov:label="process LBD"])
activity(cdr:ad, [prov:label="aggregation/disclosure protection"])
activity(cdr:asmSurvey, [prov:label="ASM Survey"])
activity(cdr:ecs, [prov:label="economic census survey"])

```

## The key PROV element embedded as DDI/XML

```

<stdyDscr> <!-- Standard DDI 2.5 -->
  <othrStdyMat> <!-- Standard DDI 2.5 -->
    <relStdy> <!-- Standard DDI 2.5 -->
      <!-- From here, PROV additions -->
      <prov:wasDerivedFrom>
        <prov:generatedEntity prov:ref="cdr:LBD"/>
        <prov:usedEntity prov:ref="cdr:BR"/>
      </prov:wasDerivedFrom>
      <prov:wasAssociatedWith>
        <prov:activity prov:ref="cdr:procLBD"/>
        <prov:agent prov:ref="cdr:CES"/>
        <prov:plan prov:ref="cdr:procLBDPlan"/>
      </prov:wasAssociatedWith>
    </relStdy> <!-- Standard DDI 2.5 -->
  </othrStdyMat> <!-- Standard DDI 2.5 -->
</stdyDscr><!-- Standard DDI 2.5 -->

```



## Additional PROV elements

These could be derived from existing DDI elements (still being developed)

```

<!-- Entities -->
<prov:entity prov:id="cdr:BR">
  <dct:title>Business Register</dct:title>
</prov:entity>
<!-- Plans = Methodology -->
<prov:plan prov:id="cdr:procLBDPlan">
  <prov:location
    xsi:type="xsd:anyURI">
    http://ideas.repec.org/p/cen/wpaper/02-17.html
  </prov:location>
  <prov:type>prov:Plan</prov:type>
</prov:plan>

```

# Work on PROV

## More details forthcoming

- ▶ Lagoze, Williams, Vilhuber “Encoding Provenance Metadata for Social Science Datasets”, submitted to Metadata and Semantics Research Conference (November 2013)
- ▶ Lagoze, Williams, Vilhuber, Block “Encoding Provenance of Social Science Data: Integrating PROV with DDI”, accepted for 5th Annual European DDI User Conference (December 2013)

# Usage scenario

## CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

[Search](#)
[Browse](#)
[Documentation](#)
[About](#)

### Filter by Codebook

IPUMSUSA

[\[info\]](#)

Longitudinal Business Database

[\[info\]](#)

NBER-CES-NAICS

[\[info\]](#)

NBER-CES-SIC

[\[info\]](#)

National QW1

[\[info\]](#)

SIPP Synthetic Beta

[\[info\]](#)

Synthetic Longitudinal Business Database

[\[info\]](#)

### Compare Variables

No variables selected

## Search

Searching Longitudinal Business Database, National QW1, Synthetic Longitudinal Business Database

[Advanced Search](#)

Show  variables

# Usage scenario

CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

Search

Browse

Documentation

About

Filter by Codebook

IPUMSUSA

[info]

Longitudinal Business Database

[info]

NBER-CES-NAICS

[info]

NBER-CES-SIC

[info]

National QWI

[info]

SIPP Synthetic Beta

[info]

Synthetic Longitudinal Business Database

[info]

Compare Variables

No variables selected

Search

Searching Longitudinal Business Database, National QWI, Synthetic Longitudinal Business Database

employment

Search

Advanced Search

Show 10 variables

Variable Name	Label	Codebook
<input type="checkbox"/> emp	synthetic March 12 Employment	Synthetic Longitudinal Business Database
<input type="checkbox"/> emp	March 12 Employment	Longitudinal Business Database
<input type="checkbox"/> qwi_eb2	QWI average employment EB2	National QWI
<input type="checkbox"/> qwi_f	QWI FQ employment	National QWI
<input type="checkbox"/> qwi_f2	QWI average FQ employment FF12	National QWI

5 variables found, displaying all variables.

Vilhuber, Abowd, Block, Lagoze, Williams

Data Management of Confidential Data

# Usage scenario

## CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

[Search](#)
[Browse](#)
[Documentation](#)
[About](#)

### Filter by Codebook

IPUMSUSA [\[info\]](#)

**Longitudinal Business Database** [\[info\]](#)

NBER-CES-NAICS [\[info\]](#)

NBER-CES-SIC [\[info\]](#)

**National QWI** [\[info\]](#)

SIPP Synthetic Beta [\[info\]](#)

**Synthetic Longitudinal Business Database** [\[info\]](#)

### Compare Variables

emp [×](#)

emp [×](#)

## Search

Searching Longitudinal Business Database, National QWI, Synthetic Longitudinal Business Database

employment

[Advanced Search](#)

Show  variables

Variable Name	Label	Codebook
<input checked="" type="checkbox"/> emp	synthetic March 12 Employment	Synthetic Longitudinal Business Database
<input checked="" type="checkbox"/> emp	March 12 Employment	Longitudinal Business Database
<input type="checkbox"/> qwi_eb2	QWI average employment EB2	National QWI
<input checked="" type="checkbox"/> qwi_f	QWI FQ employment	National QWI
<input type="checkbox"/> qwi_f2	QWI average FQ employment FF12	National QWI

5 variables found, displaying all variables.

# Usage scenario

## CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

[Search](#)
[Browse](#)
[Documentation](#)
[About](#)

Name	Label	Description	Concept	Codebook
<a href="#">emp</a>	March 12 Employment			<a href="#">Longitudinal Business Database</a>
<a href="#">qwi_f</a>	QWI: FQ employment			<a href="#">National QWI</a>
<a href="#">emp</a>	(synthetic) March 12 Employment	Paid employment consists of full and part-time employees, including salaried officers and executives of corporations, who we... <a href="#">more</a>		<a href="#">Synthetic Longitudinal Business Database</a>

© 2013 Cornell University, All Rights Reserved







Questions? Suggestions? [Email us](#)

# Highlighting provenance

## CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

[Search](#)
[Browse](#)
[Documentation](#)
[About](#)

Name	Label	Description	Concept	Codebook	Proximity
<a href="#">emp</a>	March 12 Employment			<a href="#">Longitudinal Business Database</a>	 SynLBD  National QWI
<a href="#">qwi_f</a>	QWI: FQ employment			<a href="#">National QWI</a>	 LBD  SynLBD
<a href="#">emp</a>	(synthetic) March 12 Employment	Paid employment consists of full and part-time employees, including salaried officers and executives of corporations, who we... <a href="#">more</a>		<a href="#">Synthetic Longitudinal Business Database</a>	 LBD  National QWI

© 2013 Cornell University, All Rights Reserved

Questions? Suggestions? [Email us](#)

# CED<sup>2</sup>AR next steps

- Formalize the DDI extension



# CED<sup>2</sup>AR next steps

- ▶ Formalize the DDI extension
- ▶ Provide implementation outside of Census Bureau

# CED<sup>2</sup>AR next steps

- ▶ Formalize the DDI extension
- ▶ Provide implementation outside of Census Bureau
- ▶ Test implementation within the Census RDC

# The end

## Thank you

- ▶ [3] for more details
- ▶ Labor Dynamics Institute
- ▶ VirtualRDC @ Cornell
- ▶ NCRN Cornell website

\$Id: Presentation-FCSM2013-subdoc.tex 405 2013-11-0

## Extra slides

Census Bureau

IRS

BLS

CDER

CRDC

France

# Dataset usage in Census RDC

1,505 project-dataset pairs

Many projects use multiple datasets.

# Economic (business) datasets

- ▶ 71% of datasets are business (economic) datasets
- ▶ Primarily establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD)
- ▶ They form the core of the modern industrial organization studies [5, 11] as well as modern gross job creation and destruction in macroeconomics [4, 6].
- ▶ But there are no public-use micro-data for these establishment-based products
- ▶ Exception: recently-released Synthetic LBD [2, 7]
- ▶ Currently no active curation (of derived datasets) [a], no way to reference [b], convoluted way to learn about the data structure [c\*]

# LEHD data

## Linked employer-employee data

- ▶ Longitudinal and cross-sectional detail
- ▶ New confidentiality protection methodologies [1, 10] have unlocked large amounts of data for public-use: highly detailed local area tabulations exist based on the LEHD data
- ▶ But: no public-use micro-data exist for this longitudinal job frame or any of its derivative files.
- ▶ Confidential data are dynamic (quarterly changes)
- ▶ Currently some active curation (archiving, 10-yr!) [a\*], no way to reference (publicly) [b\*], convoluted way to learn about the data structure [c\*]



# Not unique to Census Bureau

## Internal Revenue Service/ Social Security Administration

- ▶ New projects (Chetty et al, 2012; von Wachter and co-authors) have created and/or used linked longitudinal data at the IRS or the Social Security Administration.
- ▶ Neither agency has long-run experience at the statistical data curation function [a], (meta)data dissemination [b,c].
- ▶ Although both IRS and SSA have produced statistical tables for a long time.

# Not unique to Census Bureau

## Bureau of Labor Statistics

- ▶ Long history of making time-series available
- ▶ Limited access to microdata at the BLS
- ▶ Unknown curation [a]
- ▶ Even for public-use data, no way to reference specific releases [b]
- ▶ No well-established way to learn about microdata [c]

# Canadian Centre for Data Development and Economic Research



Statistics  
Canada

Statistique  
Canada

Canada



Statistics Canada

www.statcan.gc.ca

Français

Home

Contact Us

Help

Search

canada.gc.ca

[Home](#) > [The Canadian Centre for Data Development and Economic Research \(CDER\)](#) >

## The Canadian Centre for Data Development and Economic Research

Application process and guidelines

Proposal requirements

Application for accreditation

Data sets

Pricing policy

Microdata research contract

Frequently asked questions

Contact information

## Data Sets

A number of business micro databases can be accessed at CDER. Key databases are listed below. For more documentation on each of the databases, or documentation on other databases, please [contact CDER](#) at [cdcr@statcan.gc.ca](mailto:cdcr@statcan.gc.ca).

[Annual Survey of Manufacturing](#)

[Annual Survey of Manufacturing – Export and Import Registry Database](#)

[Canada Border Service Agency Customs Database](#)

[Capital and Investment Program](#)

[Longitudinal Employment Analysis Program](#)

[Longitudinal Worker File](#)

[National Accounts Longitudinal Microdata File](#)

[T2-LEAP](#)

[T2-LEAP-Export and Import Registry Database](#)

[Survey of Financing of Small and Medium Enterprises](#)

[Survey of Innovation and Business Strategies](#)

[Workplace Employee Survey](#)

## Annual Survey of Manufactures (ASM)

The ASM is a survey that covers all manufacturing locations together with associated head offices, sales offices and auxiliary units which have been classified to the manufacturing industries. Details

# Canadian Research Data Centres

RDC projects and  
publications

Conferences

FAQ

top banner, then select the "Advanced Search" option and in the field "Include pages with all these words" type in the text url:rdc and add any key word. For example, "url:rdc census" which will result in all pages on the Research Data Centres Program website that contain the keyword "census".

## Surveys available in the RDCs

The following data sets are currently available at the RDCs. For additional sources of data please refer to Statistics Canada [Products and Services](#).

To read a short **description** about a specific survey used at the RDCs, click on the survey details.

To access **detailed documentation** on a specific survey used at the RDCs, click on the appropriate cycle or year. Many of the surveys below have multiple cycles. The links below will take you to the most recent cycle or wave released. Please select "Other reference period" in the "Definitions, Data Sources and Methods Pages" for links to documentation for the earlier cycles.

Record Number	Survey Name	Acronym
5108	<a href="#">Aboriginal Children's Survey</a>	ACS
3250	<a href="#">Aboriginal Peoples Survey</a>	APS
3879	<a href="#">Adult Education and Training Survey</a>	AETS
3207	<a href="#">Canadian Cancer Registry</a>	CCR
3226	<a href="#">Canadian Community Health Survey - Annual Component</a>	CCHS
5015	<a href="#">Canadian Community Health Survey - Mental Health</a>	CCHS
5049	<a href="#">Canadian Community Health Survey - Nutrition</a>	CCHS
5146	<a href="#">Canadian Community Health Survey - Healthy Aging</a>	CCHS
5071	<a href="#">Canadian Health Measures Survey</a> <a href="#">Biobank</a>	CHMS
4440	<a href="#">Canadian Tobacco Use Monitoring Survey</a>	CTUMS
	<a href="#">Census of Population</a> - <a href="#">Additional documentation</a>	
4508	<a href="#">Ethnic Diversity Survey</a> - <a href="#">User Guide</a> - <a href="#">Codebook</a>	EDS
3504	<a href="#">Survey of Family Experiences</a>	EFES

# Canadian Research Data Centres

... but also not perfect

## Attempt to access data information on General Social Survey

**Access forbidden! / Accès interdit !**

**Access forbidden DLI!**

This web module may only be accessed from the institutional networks of Canadian postsecondary institutions participating in the Data Liberation Initiative (DLI). If you are a student or a member of a participating institution and you are unable to access these pages through your institutional network, please inform the [DLI contact at your institution](#).

**Accès interdit IDD !**

L'accès à ce module Web est restreint aux réseaux institutionnels des établissements postsecondaires canadiens membres de l'Initiative de démocratisation des données (IDD). Si vous êtes un étudiant ou personnel d'un établissement membre de l'IDD et vous ne réussissez pas à accéder à ce module par le biais de votre réseau institutionnel, veuillez informer [la personne-ressource de l'IDD à votre établissement](#).

# Réseau Quetelet

[Français](#) | [Recherche simple](#) | [Recherche avancée](#) | [Liste des enquêtes](#) | [Aide](#) | [Préférences](#) | [À propos](#) | [Votre sélection \(0\)](#)

## Réseau Quetelet

☐ français ☐ anglais  
☐ question ☐ modalités ☒ variable **revenu**

**Producteur**

Afficher 5 Filtre

<input type="checkbox"/> INSEE	984
<input type="checkbox"/> Ministère de la Santé (DREES)	34
<input type="checkbox"/> IRDES	29
<input type="checkbox"/> CEVIPOF	16
<input type="checkbox"/> Académie des Sciences Morales et Politiques - Institut de France - Fondation Simone et Cino del Duca	12

Producteurs 1 à 5 de 12 <préc. 1 3 suiv.>

**Série d'enquêtes**

Afficher 5 Filtre

<input type="checkbox"/> Enquêtes Permanentes sur les Conditions de Vie des ménages (EPCV)	295
<input type="checkbox"/> Statistiques sur les ressources et conditions de vie (SRCV)	250
<input type="checkbox"/> Enquêtes de Conjoncture Auprès des Ménages - mensuelles (ECAMME)	155
<input type="checkbox"/> Enquêtes Logement	116
<input type="checkbox"/> - Enquêtes sans série	43

Séries 1 à 5 de 21 <préc. 1 5 suiv.>

**Enquête**

Résultats 1 à 10 sur un total de 1107 pour **revenu**

Trier par **score de pertinence** Afficher 5 modalités

Question ARG - Dans le mois qui vient de s'écouler, quelles ont été vos sources de revenus ? (Plusieurs réponses possibles)

(12) 1. Vos revenus professionnels - [ARG1 - Type de revenu \(revenus professionnels de la personne\)](#)

2. Les revenus professionnels perçus par un autre membre du ménage - [ARG2 - Type de revenu \(revenus professionnels perçus par un autre membre du ménage\)](#)

3. Des pensions de retraite et préretraités - [ARG3 - Type de revenu \(pension de retraite et préretraités\)](#)

4. Pensions alimentaires - [ARG4 - Type de revenu \(pensions alimentaires\)](#)

5. Des allocations chômage - [ARG5 - Type de revenu \(allocations chômage\)](#)

Modalités (2)

Enquête [Information et Vie Quotidienne - 2004](#) - INSEE

Pertinence

Question Percevez-vous actuellement (ou votre famille perçoit-elle pour vous) une allocation, pension, ou autre revenu en raison de vos problèmes de santé ? Si oui, Lesquels ?

(14) 01. Allocation aux Adultes Handicapés (AAH) ? - [RAAH - Revenu perçu en raison de problèmes de santé : Allocation aux Adultes Handicapés](#)

02. Allocation compensatrice ? - [RACTP - Revenu perçu en raison de problèmes de santé : Allocation compensatrice](#)



J. M. Abowd, K. Gittings, K. L. McKinney, B. E. Stephens, L. Vilhuber, and S. Woodcock, "Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series," Federal Committee on Statistical Methodology, Tech. Rep., January 2012. [Online]. Available: <http://www.fcsm.gov/events/papers2012.html>



J. M. Abowd and L. Vilhuber. (2010) Synthetic data server. [Online]. Available: <http://www.vrdoc.cornell.edu/sds/>



J. M. Abowd, L. Vilhuber, and W. Block, "A proposed solution to the archiving and curation of confidential scientific inputs," in *Privacy in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer and I. Tinnirello, Eds., vol. 7556. Springer, 2012, pp. 216–225. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-33627-0\\_17](http://dx.doi.org/10.1007/978-3-642-33627-0_17)



S. J. Davis, J. C. Haltiwanger, and S. Schuh, *Job creation and destruction*. Cambridge, MA: MIT Press, 1996.



T. Dunne, M. J. Roberts, and L. Samuelson, "The Growth and Failure of U.S. Manufacturing Plants," *Quarterly Journal of Economics*, vol. 104, no. 4, pp. 671–698, 1989.



J. Haltiwanger, R. S. Jarmin, and J. Miranda, "Who creates jobs? Small vs. large vs. young," Center for Economic Studies, U.S. Census Bureau, Working Papers 10-17, Aug. 2010. [Online]. Available: <http://ideas.repec.org/p/cen/wpaper/10-17.html>



S. K. Kinney, J. P. Reiter, A. P. Reznick, J. Miranda, R. S. Jarmin, and J. M. Abowd, "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, vol. 79, no. 3, pp. 362–384, December 2011. [Online]. Available: <http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html>



C. Lagoze, W. C. Block, J. Williams, J. M. Abowd, and L. Vilhuber, "Data management of confidential data," *International Journal of Digital Curation*, vol. 8, no. 1, pp. 265–278, 2013, presented at 8th International Digital Curation Conference 2013, Amsterdam. See also <http://hdl.handle.net/1813/30924>.



C. Lagoze, W. C. Block, J. Williams, and L. Vilhuber, "Encoding provenance of social science data: Integrating prov with ddi," in *5th Annual European DDI User Conference*, accepted.



Vilhuber, Abowd, Block, Lagoze, Williams  
A. Machanavajirala, D. Kifer, J. M. Abowd, J. Gehrmann